IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 November 2022, accepted 23 December 2022, date of publication 5 January 2023, date of current version 11 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3234745

RESEARCH ARTICLE

SA-PatchCore: Anomaly Detection in Dataset With Co-Occurrence Relationships Using Self-Attention

KENGO ISHIDA[1], YUKI TAKENA[2], YOSHIKI NOTA[3], RINPEI MOCHIZUKI[3], ITARU MATSUMURA[4], AND GOSUKE OHASHI[1], (Member, IEEE)

[1]Department of Electrical and Electronic Engineering, Shizuoka University, Hamamatsu, Shizuoka 432-8561, Japan
[2]Panasonic ITS Company Ltd., Yokohama, Kanagawa 224-0054, Japan
[3]Meidensha Corporation, Numazu, Shizuoka 410-8588, Japan
[4]Railway Technical Research Institute, Tokyo 185-8540, Japan

Corresponding author: Kengo Ishida (ishida.kengo.18@shizuoka.ac.jp)

ABSTRACT Various unsupervised anomaly detection methods using deep learning have recently been proposed, and the accuracy of the anomaly detection technique for local anomalies has been improved. However, no anomaly detection dataset includes co-occurrence-related anomalies, which are combination-related. Thus, the accuracy of anomaly detection for co-occurrence-related anomalies has not progressed. Therefore, we propose SA-PatchCore, which introduces self-attention to the state-of-the-art local anomaly detection model, PatchCore. It detects anomalies in co-occurrence relationships and anomalies in local areas with the benefit of the self-attention module, which can consider contexts between separated words introduced first in the natural language processing field. As no anomaly detection dataset includes anomalies in co-occurrence relation, we prepared a new dataset called the Co-occurrence Anomaly Detection Screw Dataset (CAD-SD). Furthermore, we performed experiments on anomaly detection using the new dataset. SA-PatchCore achieves high anomaly detection performance compared with PatchCore in the CAD-SD. Moreover, our proposed model shows almost the same anomaly detection performance as PatchCore in an MVTec Anomaly Detection dataset, which is composed of anomalies in a local area. As a contribution to the anomaly detection task, we have released the CAD-SD to the public. The code and dataset are publicly available at https://github.com/IshidaKengo/SA-PatchCore

INDEX TERMS Anomaly detection, deep learning, self-attention.

I. INTRODUCTION

An anomaly detection task that identifies a sample as normal or anomalous is essential in various fields, such as industry, medical care, and security. In the industrial field, visual inspection has been conducted until now for the quality assurance of products. However, human visual inspection has problems, such as a shortage of inspectors' workforce and individual variability. Therefore, automation of appearance inspection using image recognition is expected to alleviate these problems. In recent years, deep learning has achieved outstanding results in image recognition, and various anomaly detection models using deep learning have been actively studied. The MVTec Anomaly Detection (MVTecAD) dataset [1] is used as a benchmark of deep learning-based anomaly detection techniques. The dataset is created by assuming visual inspection of products in real environments. MVTecAD [1] includes images of 15 categories of products of normal and abnormal images. The types of anomalies in the dataset are local anomalies, such as scratches, stains, and cracks, where part of the image is anomalous. Among various anomaly detection methods, the state-of-the-art PatchCore [2] achieves area under the receiver operator curve (AUROC) score of 99.6%. Many of the highly accurate methods for MVTecAD [1] use convolutional neural networks (CNNs) pre-trained using ImageNet [3] to extract features of images and distinguish normal and anomalies based on the distribution of these

The associate editor coordinating the review of this manuscript and approving it for publication was Oguzhan Urhan.

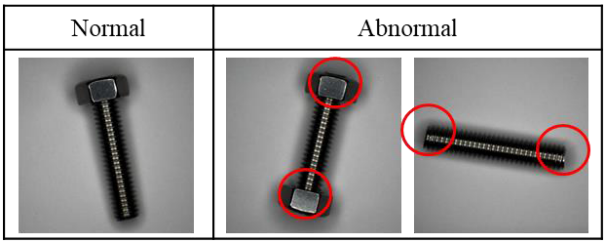

FIGURE 1. Examples of co-occurrence anomalies.

features in feature space. However, the existing detection models for MVTecAD are unable to detect anomalies in the relationships between distant pixels, which are anomalies in co-occurrence relationships because they extract image features from convolutional layers. The co-occurrence relationship anomaly is excluded from MVTecAD [1] and it is determined based on the features of the relationship between distant pixels (Fig. 1). If a product with a hex nut attached to one side of the screw rod is assumed to be normal, then it will become abnormal if a hex nut is attached to both ends of the screw rod or if there is no hex nut attached to either side of the screw rod. For such co-occurrence relation anomalies, the high-precision anomaly detection model proposed for MVTecAD [1], such as PatchCore [2], cannot sufficiently demonstrate the anomaly detection performance.

Thus, we focus on using self-attention in image recognition and enabling anomaly detection of the co-occurrence relationship. The self-attention was proposed as an operation method that can consider the relationship between words in the translating task of natural language processing [4]. Recently, there have been an increasing number of applications in the image recognition field, such as the Vision transformer [5]. We can consider the relationship between distant pixels on the image as the self-attention uses the entire image as an input and calculates the features based on the relationship between pixels. We constructed an anomaly detection method that can detect anomalies in co-occurrence relationships by capturing the relationship between distant features using self-attention. In this study, we propose a SA-PatchCore that incorporates the self-attention into PatchCore [2], which is a state-of-the-art model for MVTecAD [1], to identify anomalies in co-occurrence relationships (Fig. 2). The proposed model is valid for both anomalies in local regions and those in co-occurrence relationships. In SA-PatchCore, the local features extracted using a pre-trained CNN and the global features based on the relationship between distant pixels, obtained using self-attention to the features, are mapped on the feature space, and normal or abnormal data is distinguished based on the distribution of the features. The contribution of this study is as follows:

1) We propose SA-PatchCore incorporating self-attention into PatchCore [2] to detect anomalies in local regions and co-occurrence relationships.
2) SA-PatchCore can calculate relationships of the features without using the linear transformation and its training, which is included in the conventional self-attention model.
3) SA-PatchCore applies self-attention to compressed feature maps using the CNN so that the large computational complexity of the self-attention model does not become a bottleneck.
4) We constructed a new dataset called the Co-occurrence Anomaly Detection Screw Dataset (CAD-SD) for anomaly detection, including anomalies in the local regions and co-occurrence relationships.
5) SA-PatchCore achieves almost the same abnormality detection accuracy as PatchCore [2] for MVTecAD [1] consisting of only the abnormality in the local area while achieving a high abnormality detection performance even in the CAD-SD.

II. RELATED WORKS

A. ANOMALY DETECION USING DEEP LEARNING

In recent years, anomaly detection methods using deep learning have been significantly divided into two methods. The first is a reconstruction-based method that uses a generative model to detect abnormalities based on reconstruction errors when input images are rebuilt [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. The second is a representation-based method for detecting anomalies based on the distribution of encoded features obtained when images are put into a neural network [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35].

1) RECONSTRUCTION-BASED METHOD

The reconstruction-based method is based on generative models, such as autoencoder [17] and generative adversarial network (gan) [18]. these techniques are based on the hypothesis that the generation model learned so that only normal images can be reconstructed are unable to properly reconstruct abnormal areas of abnormal images. in the simplest case based on the autoencoder, Zhou et al. [6] performed anomaly detection by comparing input and output of Autoencoder. Bergman et al. [7] proposed an ae-ssim that replaces the error with ssim. draem [8], smai [9], and nsa [10] created pseudo-anomaly images and used them for self-supervised learning. in the gan-based methods, schlegl et al. [11] detected anomalies by comparing evaluation and generated images, and Song et al. [16] proposed anoseg using self-supervised learning. in recent years, most of the high-performance anomaly detection methods are representation-based rather than reconstruction-based methods. this is because the improved generative model successfully reconstructs abnormal images, and the methods using self-supervised learning, which uses pseudo-images, are biased against pseudo-anomalies.

2) REPRESENTATION-BASED METHOD

The representation-based method detects anomalies based on the distribution of encoded features obtained from putting images into a network. It includes the methods [19], [20],

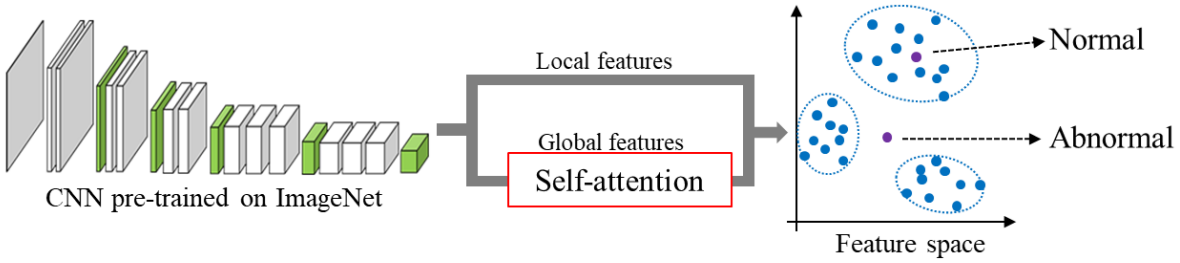

FIGURE 2. Overview of the proposed SA-PatchCore.

[21] for training a neural network to make statistical reasoning based on one-class classifications, methods [22], [23] for using the latent variable space of an autoencoder, and methods [25], [26], [27] using the discriminator of GAN to classify anomalies. However, in recent years, several methods have employed the

CNN pre-trained on large-scale external datasets, such as ImageNet, to extract image features. Different [30], CS-Flow [31], and FastFlow [32] are the representation-based methods that use the normalizing flow. SPADE [33] uses feature maps at various levels of the network for fine-grained anomaly detection and localization based on the k-NN method. The model of Rippel et al. [34] uses encoded features as a multivariate gaussian distribution and calculates anomaly scores using the Mahalanobis distance. PaDiM [35] applies this approach at the patch-level to multi-scale feature maps. Several structures of the SPADE and PaDiM are related to PatchCore [2], which is the current state-of-the-art anomaly detection model in MVTecAD benchmark [1].

PatchCore [2] uses the Wide-ResNet50 [36], pre-trained on the ImageNet, as a feature extractor and average pooling to aggregate feature maps extracted from the middle layer of Wide-ResNet50 [36] to calculate the features per patch. The features of the calculated normal data are stored in the memory bank during training. Furthermore, the features of the calculated unknown data and the feature quantity in the memory bank with a small distance on the feature space are obtained using the k-NN method during inference. The distance is used as the patch-level anomaly score and the maximum of this patch-level anomaly score is the image-level anomaly score. PatchCore [2] reduces the loss of normal and abnormal information by considering the neighbor pixels for patch-level features. Greedy Coreset Subsampling reduces computational costs. PatchCore [2] can detect anomalies with high accuracy for anomalies in local areas in datasets, such as MVTecAD [1]. However, PatchCore [2] is weak to anomalies in co-occurrence relationships because it is the mechanism for extracting features using a pre-trained CNN. The proposed SA-PatchCore solves this PatchCore problem [2] by applying the self-attention to the extracted features, and it can detect the anomalies of co-occurrence relationships.

B. SELF-ATTENTION

Self-attention is proposed for natural language processing translating tasks [4], which can consider the context between distant words. Specifically, the input sequence is linearly transformed to generate three variables: query, key, and value. The inner product of the query and key is normalized using softmax to obtain the relevance of the key (search destination) to the query (search source). The weighted sum of this relevance and the value is the output of the self-attention. Therefore, the self-attention module, which can consider the relevance of the entire input sequence, solved the problem of relevance disappearing because of the distance of the input sequence of recurrent neural networks used in the conventional machine translation. The module achieved model features based on the global feature relevance in the input image regardless of the distance of the input sequence.

In recent years, using self-attention has been actively studied, even in the image recognition field. SASA [37], LRNet [38], SANet [39], and Axial-SASA [40] proposed a model, in which the self-attention layer replaces the convolution layer in ResNet, as a simple approach to use self-attention in image recognition. Each of these models proposes to replace self-attention in a different format. The Vision Transformer [5] proposes a model structure that divides the input images into patches and puts these patches into several transformer block. It shows comparable performance to or better than the conventional CNN. DETR [41], VideoBERT [42], VILBERT [43], CCNet [44], AA-CN [45], and BoTNet [46] are models using both convolution and self-attention. The computational complexity becomes enormous when high-resolution images are input into self-attention because its computational complexity increases in the order of square based on the length of the input sequences. BoTNet [46] applies self-attention to feature maps whose resolution is reduced using convolution to solve this problem. Furthermore, our proposed SA-PatchCore has a similar construction and prevents the computational complexity from increasing because it uses self-attention for feature maps compressed using a pre-trained CNN.

III. METHOD

Our proposed model is based on PatchCore [2], which is a state-of-the-art anomaly detection model in MVTecAD [1] and introduces the self-attention module. We named the proposed model SA-PatchCore. SA-PatchCore retains the high anomaly detection performance of PatchCore [2] for local anomalies, and the introduction of the self-attention module

enables highly accurate anomaly detection in co-occurrence relationships. Fig. 3 depicts the model structure of SA-PatchCore.

A. PatchCore-BASED STRUCTURE

SA-PatchCore is based on PatchCore [2] and is composed of several parts.

1) FEATURE EXTRACTION

SA-PatchCore uses the WideResNet50 [36] pre-trained on ImageNet to extract features of input images. The final output of each hierarchy from the convolutional network is extracted as a feature map and used for abnormality detection. Generally, the deeper the hierarchy, the more the global feature map captured, which is specialized for learning tasks. SA-PatchCore uses feature maps of the middle layers of the WideResNet50 [36] because the local features for the unknown data are crucial in the industrial anomaly detection task. Specifically, SA-PatchCore uses Layers 2 and 3 of the WideResNet50 [36]. Layer 2 has a more local feature representation than Layer 3; the algorithm of PatchCore that aggregates features in the neighborhood is applied to Layer 2 to detect local anomalies. Let $\phi_2(h, w, c)$ be the feature map of Layer 2 with height h, width w, and c channels. The patch-level features that aggregate local features in the neighborhood are expressed as follows:

$$P_2 = f_{agg}(\phi_2) \tag{1}$$

f_{agg} is the aggregate function in the neighborhood. SA- PatchCore [2] uses average pooling with a kernel size of 3, stride 1, and padding 1. As Layer 3 has a more global feature representation than Layer 2, its feature is used as input to the self-attention module for detecting anomalies in co-occurrence relationships. Let the feature map of Layer 3 be $\phi_3(h, w, c)$ and the self-attention module be a transformation function f_{SA} to features with information necessary for anomaly detection of co-occurrence relationships. The features considering relationships obtained from Layer 3 are expressed as follows:

$$P_3 = f_{SA}(\phi_3) \tag{2}$$

P_2, which aggregates features in the neighborhood to detect local anomalies, and P_3, which contains the information necessary for anomaly detection of co-occurrence relationships, are concatenated and stored in a memory bank M. The resolution of P_3 resized to match that of P_2 since it has a lower resolution thanP_2.

2) CORESET SUBSAMPLING

The size of the required memory bank becomes large and inference time significantly increases when the size of the feature map increases. PatchCore [2] solves this problem by subsampling the feature quantity using greedy coreset subsampling, and SA-PatchCore uses a similar mechanism. Coreset subsampling finds a subset $S \in A$, such that the solution to the problem in sample A comes closest to that of sample S [47]. The coreset $M_c$for the memory bank M in the patch-level feature space is chosen so that the coverage of M_c is approximately the same as the original memory bank M [48], [49] because PatchCore [2] takes the nearest neighbor computation. PatchCore [2] uses the iterative greedy approximation proposed in [49] because the exact computation of M_c is NP-hard.

TABLE 1. Shooting environment of Co-occurrence Anomaly Screw dataset.

Camera	DFK33UX183 (Argo Co.)
Camera Aperture	16 mm
Shooting Distance	25 cm
Shooting Image Size	5472×3648
Image Size	700×700
Lighting	HPR2-75SW (CCS Co.)
Power source for lightning	PD2-5024(A)

3) ANOMALY DETECTION

SA-PatchCore selects m^* which is the nearest neighbor of the patch-level features $m^{test}$of test data, among the patch-level features $m \in M$ of the training data stored in the memory bank. It estimates the patch-level anomaly score s of the test image X^{test} from the distance between patch-level features m^{test} and m^*.

$$\begin{aligned} m^* &= argmin_{m \in M} \left\| m^{test} - m \right\|_2 \\ s &= \left\| m^{test} - m^* \right\|_2 \end{aligned} \tag{3}$$

The image-level anomaly score$S$ for the test image X^{test} is obtained from the maximum patch-level anomaly score s in the X^{test}.

B. SELF-ATTENTION MODULE

SA-PatchCore introduces a self-attention module (Fig. 4) to detect co-occurrence anomalies. This module is applied to the feature maps obtained from Layer 3 of the WideResNet50 [46], and it is used as a transformation module to obtain feature maps X_{SA} with the information required for the anomaly detection of co-occurrence relationships. Once a feature map of Layer 3 $\phi_3(h, w, c)$ is obtained, max pooling of kernel sizes 3, strides 1, and padding 1 are applied to emphasize the nearby features, which are turned into vectors $X \in R^{hw \times c}$.X is replicated in triplicate to compute the self-attention as a query, key, and value in the Transformer [4]. X_{SA} is expressed as follows:

$$X_{SA} = \text{softmax}\left(\frac{XX^T}{\sqrt{d_X}}\right)X \tag{4}$$

where d_X is the depth of X. The vector X_{SA}, which considers the relationship between distant features, is obtained by calculating the relationship between pixels using the product of a query and key as weights and calculating the product of the weights multiplied by the softmax and value. By resizing the obtained X_{SA} to the size of the original feature map,

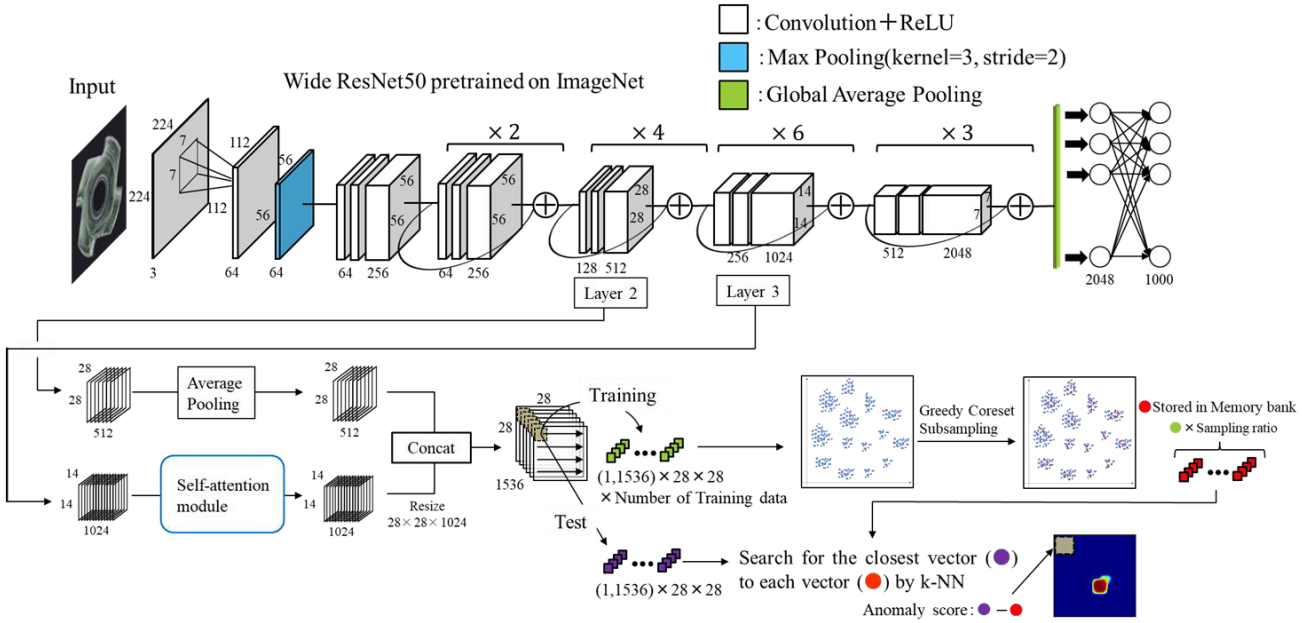

FIGURE 3. Structure of the SA-PatchCore.

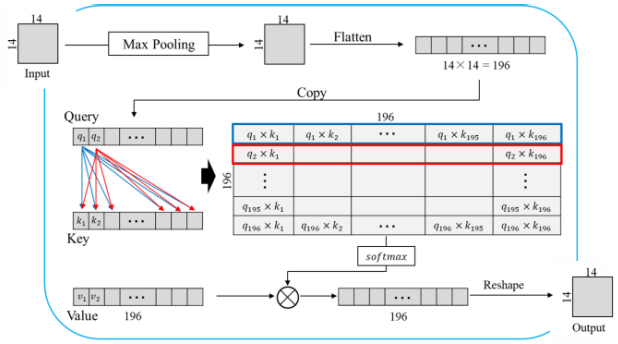

FIGURE 4. Overview of the self-attention module.

a feature map P_3 with the information necessary for detecting anomalies of co-occurrence relationships is obtained. The self-attention module does not calculate keys, queries and values by linear transformation in the Transformer [4] but uses max pooling because it is used to generate feature maps based on the relationships necessary to detect anomalies in co-occurrence relationships. Furthermore, since the computational complexity of the self-attention increases in the order of the square based on the input sequence length, the high computational complexity is occasionally a problem when high-resolution images are input into the self-attention. However, SA-PatchCore has the advantage that the computational complexity problem does not become a bottleneck because it inputs feature maps compressed using a pre-trained CNN to the self-attention module.

IV. EXPERIMENTS

We created the CAD-SD to verify the effectiveness of SA-PatchCore, which includes the anomaly of the local area and that of the co-occurrence relationship. Then, we experimented with anomaly detection on the dataset.

A. CO-OCCURRENCE ANOMALY DETECTION SCREW DATASET (CAD-SD)

MVTecAD [1] is a typical dataset for evaluating the anomaly detection method; however, it contains only the abnormality of a local area, in which an abnormal part exists only in some parts, such as scratches and dirt. Currently, there is no dataset for anomalies of co-occurrence relationships, which are anomalies of combinatorial relationships. Therefore, we created the CAD-SD, which includes the anomaly of the local area and that of the co-occurrence relationship for the images of products consisting of screw rods and hex nuts. The images in the dataset were taken at random angles using a camera. Table 1 shows the imaging environment of the dataset. The camera used was a DFK33UX183 manufactured by Argo Corporation. The aperture and shooting distance were set at 16 and 25 cm respectively. The size of the image in the dataset was trimmed from 5472 × 3648 to 700 × 700. HPR2-75SW manufactured by CCS Corporation was used for the lighting, and PD2-5024 (A) was used for the power supply. Figure 5 shows examples of the images in the dataset. The CAD-SD includes normal images of the product with a hex nut attached to one side of the screw rod. The types of abnormal images in the dataset are roughly divided into the anomalies of the local region and that of the co-occurrence relation. The anomalies of the local area are "*Scratch*," in which a portion of the product is scratched, and "*Paint*," in which some paint adheres to a part of the product. The anomalies in the co-occurrence relationship are "*Over-coupling*," where hex nuts are coupled on both sides of the screw rod, and "*Lacking*," where hex nuts are not

TABLE 2. Accuracy of anomaly detection on CAD-SD (AUROC).

Method	PatchSVDD [20]	PaDiM [35]	PatchCore [2]	CS-Flow [31]	SA-PatchCore
AUROC	73.1	65.8	83.5	92.0	**97.6**

TABLE 3. Accuracy for each anomaly category on Co-occurrence Anomaly Screw Dataset (AUROC). Red and blue stand for the first and second places respectively.

Anomaly category		PatchSVDD [20]	PaDiM [35]	PatchCore [2]	CS-Flow [31]	SA-PatchCore
Local anomaly	Scratch	74.8	79.9	**99.6**	89.8	**98.0**
	Paint	74.1	89.2	**99.8**	93.0	**99.7**
	Average	74.5	84.6	**99.7**	91.4	**98.9**
Co-occurrence anomaly	Over-coupling	72.1	68.8	89.6	**90.6**	**99.7**
	Lacking	72.1	24.2	46.5	**94.9**	**92.9**
	Average	72.1	46.5	68.1	**92.8**	**96.3**

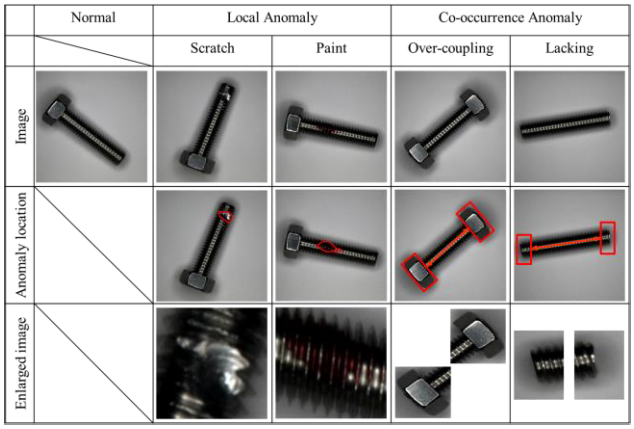

FIGURE 5. Example images of Co-occurrence anomaly screw dataset.

coupled on either side of the screw rod. There are 400 normal training images. For the evaluation, "*Normal*," "*Scratch*," "*Paint*," "*Over-coupling*," and "*Lacking*" contain 210, 41, 41, 44, and 40 images, respectively. The CAD-SD is publicly available at present.

B. EXPERIMENTAL CONDITION

We experimented with anomaly detection using the CAD-SD. The image in the dataset was resized to 224 × 224 and used as input to the model. The CPU is an Intel®Core i9-9900K CPU @ 3.60 GHz, and the memory is 32 GB. The GPU configuration is an NVIDIA GeForce RTX 3090 with 24 GB of memory. The batch size is 1 and the sampling rate of Greedy Coreset Subsampling is 1%. PatchSVDD [20], PaDiM [35], PatchCore [2], and CS-Flow [31] are used as comparison methods. The AUROC is used as the evaluation metric for image-level anomaly detection; the AUROC was calculated for all test images and each anomaly type.

C. RESULTS

Table 2 shows the results of image-level anomaly detection in CAD-SD. Table 3 shows the evaluation for each type of anomalies in CAD-SD. SA-PatchCore achieved the best performance. SA-PatchCore is slightly less accurate than PatchCore [2] in detecting anomalies in the local regions of "*Scratch*" and "*Paint*," but it is more accurate than the other methods. The method is on average about 30% more accurate than PatchCore [2] in detecting anomalies of the co-occurrence relationship between "*Over-coupling*" and "*Lacking*," which is the highest accuracy. This result indicates that SA-PatchCore has a significant improvement in the detection of co-occurrence anomalies while maintaining sufficient detection performance for anomalies in local regions. It shows the advantage that SA-PatchCore retains the effectiveness of PatchCore [2] for anomalies in local regions while improving the effectiveness for anomalies in co-occurrence relationships by introducing the Self-attention module. Figure 6 shows the results of localizing the anomaly area. The heatmap is normalized based on the patch-wise anomaly scores of all test images, and the lower limits are set to appropriate values. The red color indicates that the anomaly score is higher. SA-PatchCore is able to identify both local anomalies and co-occurrence anomalies. Table 4 shows the inference speed for a single image on CAD-SD. SA-PatchCore achieves almost the same inference speed as PatchCore [2], which is faster than the other methods. It indicates that SA-PatchCore achieves high detection accuracy by introducing self-attention while maintaining a high inference speed.

V. DISCUSSION

Several discussions are presented on SA-PatchCore. First, we evaluated the anomaly detection performance on several anomaly detection datasets including MVTecAD [1]. Next, we examined the optimization of the modeling structure by focusing on the hierarchy of feature extraction and pooling in the self-attention module.

A. ANOMALY DETECTION ON OTHER DATASETS

We experimented with MVTecAD [1] to investigate the anomaly detection performance of the SA-PatchCore, which is a widely used anomaly detection dataset, although it excludes co-occurrence anomalies. Table 5 shows that the anomaly detection performance of SA-PatchCore on MVTecAD [1] was slightly lower than that of PatchCore [2]

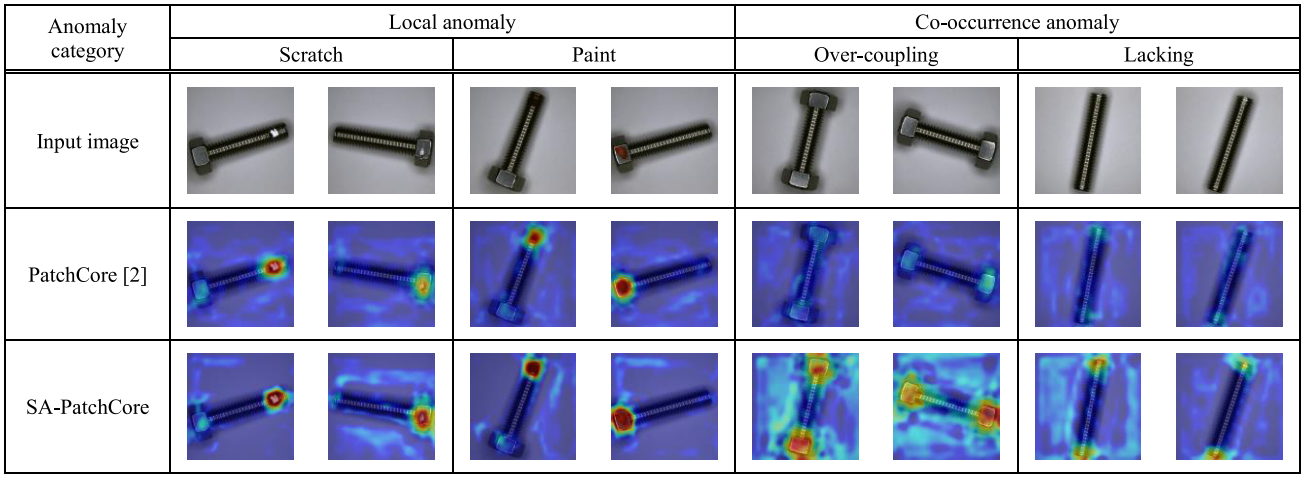

FIGURE 6. Localization of anomaly areas on CAD-SD.

TABLE 4. Mean inference time per an image on CAD-SD.

Method	PatchSVDD [20]	PaDiM [35]	PatchCore [2]	CS-Flow [31]	SA-PatchCore
Inference time (ms)	680.0	275.2	24.0	61.8	25.0

TABLE 5. Accuracy of anomaly detection on MVTecAD [1] (AUROC).

Method	PatchSVDD [20]	PaDiM [35]	PatchCore [2]	CS-Flow [31]	SA-PatchCore
AUROC	91.3	95.5	98.6	98.7	97.1

TABLE 6. Accuracy of anomaly detection on BTAD [50] and AITEX [51] (AUROC).

Dataset	PatchCore [2]	SA-PatchCore
BTAD [50]	92.3	93.7
AITEX [51]	85.8	89.0

and CS-Flow [31] but better than PatchSVDD [20] and PaDiM [35]. Table 6 shows the results on the BeanTech Anomaly Detection dataset (BTAD) [50] and the AITEX dataset [51]. SA-PatchCore scores higher detection accuracy than PatchCore [2] for these datasets. SA-PatchCore has the advantage of being able to detect both local anomalies and co-occurrence anomalies well. However, these existing datasets exclude co-occurrence anomalies and consist mainly of local anomalies. These results show that SA-PatchCore has sufficient anomaly detection performance even for datasets consisting of only local anomalies. SA-PatchCore has high anomaly detection performance even for local anomalies, while improving the anomaly detection performance of co-occurrence relations by introducing the Self-attention module.

B. OPTIMIZATION OF THE MODEL STRUCTURE

1) HIERARCHY OF FEATURE EXTRACTION

The proposed SA-PatchCore places Layer 2 of the WideResNet50 [46] into the average pooling for local feature extraction and Layer 3 into the self-attention module for feature extraction of co-occurrence relationship. To evaluate the validity of this structure, we conducted anomaly detection experiments on the CAD-SD even in a model structure where Layers 2 and 3 are combined and inputted into the average pooling and the self-attention module. This structure directly incorporates the self-attention module into PatchCore [2]. Table 7 shows that the structure of SA-PatchCore is more effective than the original structure of PatchCore [2], which uses Layers 2 and 3 cooperatively. This confirms that SA-PatchCore is a suitable model structure for detecting anomalies in local regions and co-occurrence relationships.

TABLE 7. Anomaly detection performance by hierarchy of feature extraction.

SA-PatchCore	SA-PatchCore (Layer 2 + Layer 3)
97.6	96.7

TABLE 8. Anomaly detection performance using pooling in the self-attention module.

Max Pooling	Average Pooling	Without Pooling
97.6	92.6	90.9
97.6	92.6	90.9

2) POOLING IN THE SELF-ATTENTION MODULE

We investigated the suitability of max pooling in the self-attention module for SA-PatchCore in the CAD-SD when average pooling or no pooling is used instead of max pooling. The results in Table 8 show that the anomaly detection performance is the best when max pooling is used, which is effective for detecting anomalies in co-occurrence relationships.

VI. CONCLUSION

We proposed SA-PatchCore in this study, which extends the current state-of-the-art PatchCore [2] to detect anomalies in co-occurrence relationships by introducing a self-attention module. This module is a transformation module that can obtain feature maps by considering the relationship between features without using the linear transformation of the conventional self-attention and its training. SA-PatchCore prevents the computation of self-attention from computational complexity by inputting feature maps compressed using a pre-trained CNN in the self-attention module. Furthermore, since no anomaly detection dataset includes co-occurrence anomalies, we prepared the CAD-SD that includes both local and co-occurrence anomalies. SA-PatchCore has sufficient anomaly detection performance on MVTecAD [1], which is composed of only local anomalies, and it achieves state-of-the-art anomaly detection performance in the CAD-SD.

REFERENCES

[1] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9592–9600.

[2] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," 2021, *arXiv:2106.08265*.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, and L. Kaiser, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2007, pp. 5998–6008.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[6] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 665–674.

[7] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," 2018, *arXiv:1807.02011*.

[8] V. Zavrtanik, M. Kristan, and D. Skocaj, "DRAEM—A discriminatively trained reconstruction embedding for surface anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8330–8339.

[9] Z. Li, N. Li, K. Jiang, Z. Ma, X. Wei, X. Hong, and Y. Gong, "Superpixel masking and inpainting for self-supervised anomaly detection," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2020, pp. 1–12.

[10] H. M. Schlüter, J. Tan, B. Hou, and B. Kainz, "Natural synthetic anomalies for self-supervised anomaly detection and localization," 2021, *arXiv:2109.15222*.

[11] T. Schlegl, P. Seebock, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *Proc. Int. Conf. Inf. Process. Med. Imag. (IPMI)*, 2017, pp. 146–157.

[12] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, "Efficient GAN-based anomaly detection," 2018, *arXiv:1802.06222*.

[13] S. Akcay, A. Atapour-Abarghouei, and T. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 622–637.

[14] M. Sabokrou, M. Pourreza, M. Fayyaz, R. Entezari, M. Fathy, J. Gall, and E. Adeli, "Avid: Adversarial visual irregularity detection," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 488–505.

[15] P. Perera, R. Nallapati, and B. Xiang, "OCGAN: One-class novelty detection using GANs with constrained latent representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2898–2906.

[16] J. Song, K. Kong, Y.-I. Park, S.-G. Kim, and S.-J. Kang, "AnoSeg: Anomaly segmentation network using self-supervised learning," 2021, *arXiv:2110.03396*.

[17] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[18] I. Goodfellow, J. Abadie, M. Mirza, B. Xu, D. Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014, pp. 2672–2680.

[19] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 4393–4402.

[20] J. Yi and S. Yoon, "Patch SVDD: Patch-level SVDD for anomaly detection and segmentation," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2020, pp. 375–390.

[21] F. V. Massoli, F. Falchi, A. Kantarci, S. Akti, H. K. Ekenel, and G. Amato, "MOCCA: Multilayer one-class classification for anomaly detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 6, pp. 2313–2323, Jun. 2022.

[22] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 481–490.

[23] P. Schlachter, Y. Liao, and B. Yang, "One-class feature learning using intra-class splitting," in *Proc. 27th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2019, pp. 1–5.

[24] W. Liu, R. Li, M. Zheng, S. Karanam, Z. Wu, B. Bhanu, R. J. Radke, and O. Camps, "Towards visually explaining variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8639–8648.

[25] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3379–3388.

[26] M. Z. Zaheer, J.-H. Lee, M. Astrid, and S.-I. Lee, "Old is gold: Redefining the adversarially learned one-class classifier training paradigm," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14171–14181.

[27] Z. Zhang, S. Chen, and L. Sun, "P-KDGAN: Progressive knowledge distillation with GANs for one-class novelty detection," 2020, *arXiv:2007.06963*.

[28] T. Reiss, N. Cohen, L. Bergman, and Y. Hoshen, "PANDA: Adapting pretrained features for anomaly detection and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2805–2813.

[29] P. Liznerski, L. Ruff, R. A. Vandermeulen, B. J. Franks, M. Kloft, and K.-R. Müller, "Explainable deep one-class classification," 2020, *arXiv:2007.01760*.

[30] M. Rudolph, B. Wandt, and B. Rosenhahn, "Same same but DifferNet: Semi-supervised defect detection with normalizing flows," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1906–1915.

[31] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, "Fully convolutional cross-scale-flows for image-based defect detection," 2021, *arXiv:2110.02855*.

[32] J. Yu, Y. Zheng, X. Wang, W. Li, Y. Wu, R. Zhao, and L. Wu, "FastFlow: Unsupervised anomaly detection and localization via 2D normalizing flows," 2021, *arXiv:2111.07677*.

[33] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences," 2020, *arXiv:2005.02357*.

[34] O. Rippel, P. Mertens, and D. Merhof, "Modeling the distribution of normal data in pre-trained deep features for anomaly detection," 2020, *arXiv:2005.14140*.

[35] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: A patch distribution modeling framework for anomaly detection and localization," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*. Cham, Switzerland: Springer, 2021, pp. 475–489.

[36] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*.
[37] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," 2019, *arXiv:1906.05909*.
[38] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3464–3473.
[39] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10076–10085.
[40] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation," 2020, *arXiv:2003.07853*.
[41] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020, *arXiv:2005.12872*.
[42] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7464–7473.
[43] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 13–23.
[44] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.
[45] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention augmented convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3286–3295.
[46] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16519–16529.
[47] P. Agarwal, S. Har, P. Kasturi, and R. Varadarajan, "Geometric approximation via coresets," in *Combinatorial and Computational Geometry* (Mathematical Sciences Research Institute Publications), vol. 52, no. 11. Berkeley, CA, USA, 2004.
[48] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," 2017, *arXiv:1708.00489*.
[49] S. Sinha, H. Zhang, A. Goyal, Y. Bengio, H. Larochelle, and A. Odena, "Small-GAN: Speeding up GAN training using core-sets," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 9005–9015.
[50] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, and G. L. Foresti, "VT-ADL: A vision transformer network for image anomaly detection and localization," in *Proc. IEEE 30th Int. Symp. Ind. Electron. (ISIE)*, Jun. 2021, pp. 1–6.
[51] J. Silvestre-Blanes, T. Albero-Albero, I. Miralles, R. Pérez-Llorens, and J. Moreno, "A public fabric database for defect detection methods and results," *Autex Res. J.*, vol. 19, no. 4, pp. 363–374, Dec. 2019.

KENGO ISHIDA received the B.E. degree from the Department of Electrical and Electronic Engineering, Faculty of Engineering, Shizuoka University, Hamamatsu, Japan, in 2022. He is currently enrolled at the Department of Electrical and Electronic Engineering, Graduate School of Integrated Science and Technology, Shizuoka University. His research interests include deep learning, artificial intelligent, and computer vision.

YUKI TAKENA received the B.E. degree from the Department of Electrical and Electronic Engineering, Faculty of Engineering, Shizuoka University, Hamamatsu, Japan, in 2020, and the M.E. degree from the Department of Electrical and Electronic Engineering, Graduate School of Integrated Science and Technology, Shizuoka University, in 2022. He currently works at Panasonic ITS Company.

YOSHIKI NOTA received the B.E. and M.E. degrees from Kagawa University, Takamatsu, Japan, in 2012 and 2014, respectively. He joined Meidensha Corporation, in 2014. He works research and development in the areas of image recognition and image measurement.

RINPEI MOCHIZUKI received the B.E. and M.E. degrees from Shizuoka University, Hamamatsu, Japan, in 2015 and 2017, respectively. He joined Meidensha Corporation, in 2017. He works research and development in the areas of image processing, image recognition, and anomaly detection.

ITARU MATSUMURA received the B.E. and M.S. degrees in electrical engineering from the Tokyo Institute of Technology, Tokyo, in 2007 and 2009, respectively. Since April 2009, he has been with the Railway Technical Research Institute, Tokyo.

GOSUKE OHASHI (Member, IEEE) received the B.E., M.E., and D.E. degrees from Keio University, Yokohama, Japan, in 1992, 1994, and 1997, respectively. He has been an Assistant Professor since 1997. He is currently a Professor at the Department of Electrical and Electronic Engineering, Shizuoka University. He was a Visiting Researcher at the University of California, Santa Barbara, from 2003 to 2004. His research interests include image processing, computational vision, and visual perception.